2017
THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

PDF/UA Structure Elements and the User Experience

Karen McCall, M.Ed.

Karlen Communications

info@karlencommunications.com

Abstract

There is a direct correlation between the experiences of people with disabilities and the correct implementation of PDF structure elements in PDF documents.

Keywords

Government, Information and Communication Technology, Research and Development.

Journal on Technology and Persons with Disabilities
Santiago, J. (Eds): CSUN Assistive Technology Conference

Introduction

In 2015 the *PDF and the User Experience Survey* was conducted to quantify the experiences of those with disabilities who use adaptive technology to access PDF documents (McCall). The results provided qualitative confirmation of what those of us in the field have known since Tags were introduced in 2001:

- There is a lot of inaccessible PDF content out there, either because it isn't tagged or it is tagged incorrectly.
- That many people using adaptive technology are using third-party tools to take PDF content out of the PDF format in order to access it and ensure that it is *portable*.
- That a "properly tagged PDF document is a pleasure to read."

The following are only a few comments from the 146 respondents who took part in the PDF and the User Experience Survey in 2015:

- "The viewing tools on Mac and iPhone don't allow [me] to navigate through properly tagged PDF by heading, bookmarks and so on. The structure of the document is not rendered." (Respondent 4427724923)
- "Any time I try [to read a PDF document] it usually says blank document, or if it does read it there's no tags, and when there's no tags its hard to figure out what's what. Then you've got words that are either split up or have errors, like to the point you cant figure out what it is." (Respondent 4426566325)
- "For work most PDFs are complex layouts including equations. If I try to reflow to increase the zoom level etc., then it destroys the equations. There is inadequate control of font, font size, interline spacing and layout. In a Word document I would control all of these for easier reading. Sometimes you can't search properly and this is

annoying as this is how I most easily locate information (scanning through is harder). Sometimes you copy and paste and get rubbish." (Respondent 4398401216)

These are only a few of the frustrations reported by people using adaptive technology in trying to access PDF documents, even those that are tagged.

Discussion

Structure Elements and Content

PDF documents are prevalent because they can be opened on any platform or device and they retain the integrity of layout. They are a visual representation of a physical page/document. How does this visual representation of a page/document become something that can be read and comprehended by someone using adaptive technology (PDF Association)? At the root of a tagged PDF document is a Tag called <Tags>. This is also known as the Tags Root. From the parent <Tags> root, all other structural elements of the document are nested. The relationship of Tags is defined using a parent/child analogy. Nested under the <Tags> root is the child <Document> Tag.

A PDF document can have Heading Tags. These are represented with Tags such as <H1> for a Heading level 1, <H2>, <H3> and so forth. It is essential that Headings, which denote topic changes and the structure of the document, be sequential and not skip levels. For example, a document can't skip from an <H1> to an <H3> without an <H2> between them. Headings are the primary means of navigating through a PDF document for those of us using adaptive technology. Most adaptive technology can either list Headings or move from Heading to Heading (topic to topic).

Tagged PDF documents use <P> Tags for content on a page that should be read as a paragraph. All lines in a paragraph must be under a single <P> Tag so that the lines are read as a continuous stream of information as opposed to being read as individual paragraphs.

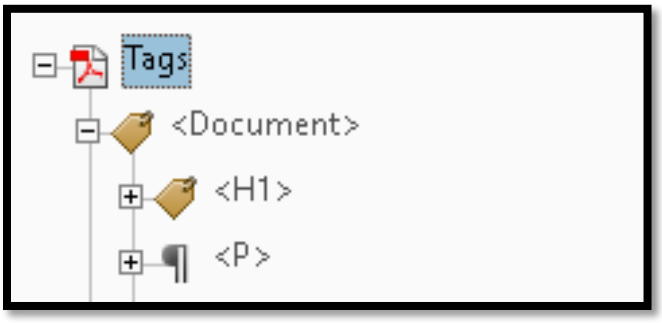

Fig. 1. Sample Tags Tree with Heading and Paragraph Tags.

Another structure element that is prevalent in the visual representation of a page is a **list**. Lists have several structure elements that signal to adaptive technology that a list is being encountered, how many items are in the list, each list item, and that the person is leaving the list and returning to "normal" text.

The list structure has a parent <L> Tag under which are child <LI> or List Item Tags. There is one <LI> Tag for each item in the list. Nested under each <LI> Tag is a child <Lbl> Tag representing a bullet or number and a child <LBody> Tag representing the text or content of the List Item. The <Lbl> Tag is optional and is dependent on whether there is a bullet or number associated with the list. As long as there is a list structure, the "content" will be read as a list of related items.

Many documents have tables. The structure elements for tables are similar to those of lists in that there are a prescribed sequence of Tags that create a table structure. The parent <Table> Tag signals that a table has been encountered. The <Table> has a series of child <TR> Tags representing Table Rows and each Table Row has either a child Tag of <TH> Tag for a Table Header (column or row title) or a child Tag of <TD> for a Table Data cell. Tables can also

have child Tags of <THead> for a Table Header (not a Heading) And a <TFoot> for a Table Footer.

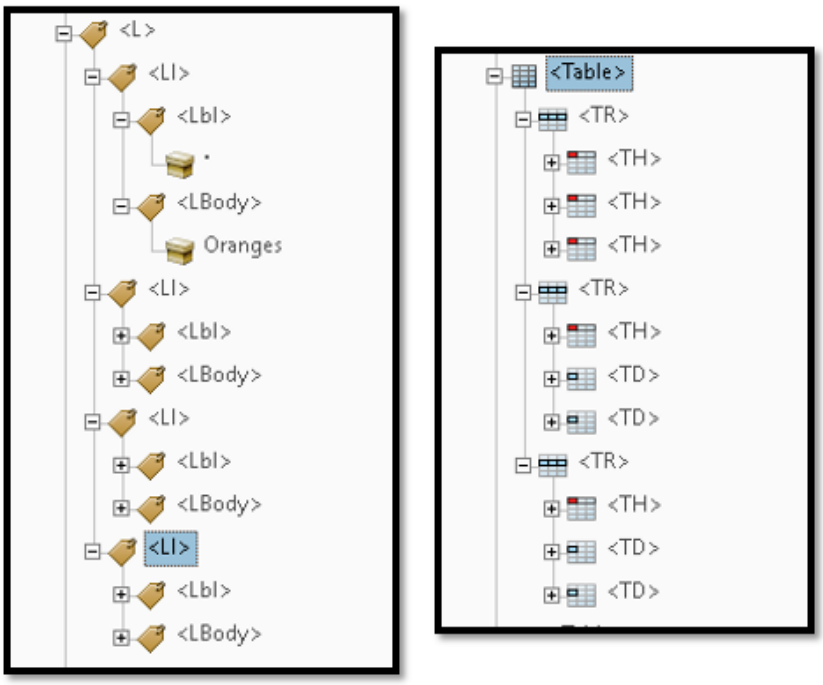

Fig. 2: Sample List structure (Left) and Table structure (right).

The correct semantic structure for accessible hyperlinks is equally important; A person must be able to activate a link using the keyboard. The <Link> Tag is used to identify links in tagged PDF documents but the <Link> Tag must have a "Link-OBJR" or *Link Object* nested under it along with the web address or text on the visual representation of the page that is the link.

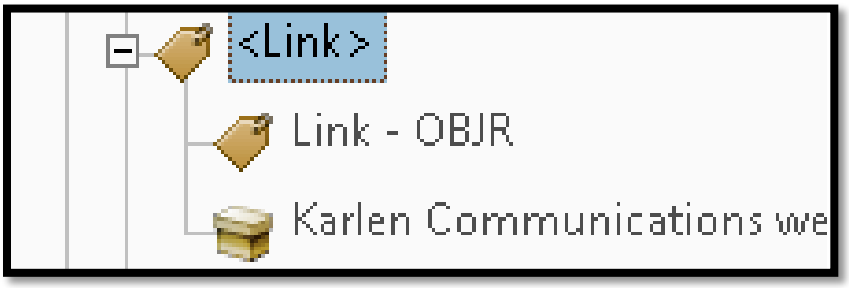

Fig. 3. Sample Link structure.

This brief summary of some basic Tags demonstrates the importance of using the correct structural elements in tagged PDF documents; Tags control how the information is rendered to the end user, and provide a method of searching and navigating the PDF document.

How Did We Get to This Point?

The ability to Tag PDF documents for accessibility began with the introduction of PDF Specification 1.4 in Adobe Systems Acrobat 5 in 2001. This capability set the stage to begin examining how we work with and create accessible digital content. The first publication to address the needs of document authors and remediators working with content that would eventually become tagged accessible PDF documents was published in October 2005 (Karen McCall, Accessible and Usable PDF Documents: Techniques for Document Authors).

In response to this publication and in an attempt to answer the question "what can we do in a native application to alleviate the amount of repairs in Acrobat", a follow-up book on creating more accessible Microsoft Word documents was published in 2005 (Karen McCall, Logical Document Structure Handbook: Word 2003). As Section 508 awareness expanded in the United States, document authors and remediators asked how non-HTML content could be mapped to the Section 508 legislation. In November 2006 one of the first attempts to map general document accessibility to Section 508 was published (Karen McCall, Mapping Section 508 to Digital Document Accessibility).

In 2004 an international working group was established to create a standard for PDF documents. The group was incorporated into the ISO (International Standards Organization) and AIIM (Association for Information and Image Management) in 2009. In 2012 the first PDF/UA or ISO 14289 international standard for accessible PDF was published (PDF Association). ISO 32000 was published in 2008 and put the PDF format in the public domain, no longer proprietary

to Adobe Systems. PDF/UA builds on chapter 14 of ISO 3200 which identifies the Tag set used to make a PDF document accessible (International Standards Organization). This paper uses examples of structural elements/Tags to demonstrate the necessity to correctly Tag PDF documents.

As we move toward PDF/UA 2, there is still confusion about syntax/structure elements, how to Tag a PDF document and a lack of tools to do so efficiently, as well as a lack of support for accessible document design and creation in applications that output digital content. There are many PDF documents that are scanned images of pages and there are PDF documents for which there is no original document to refer to. The correct tagging of a PDF document is essential in providing access to content and structure of the visual representation of a page/document (PDF Association).

Conclusion

Some beneficial by-products of the need to remediate PDF documents are: an improvement in Acrobat's remediation tools, the development of best practices by the industry, the incorporation of accessibility checkers in many native applications used to create the original source documents, such as Microsoft Word, and the incorporation of PDF/UA into revisions of legislation for digital content (United States Access Board). It is critical to the user experience that PDF documents be tagged using the correct structure elements for the visual representation of the content on the page. We need the tools in all applications to make this easier and efficient, reducing labour and time costs and avoiding the purchase of expensive tools as add-ons to existing expensive ones.

Works Cited

International Standards Organization. *ISO 14289-1:2012 , Document management applications - - Electronic document file format enhancement for accessibility - Part 1: Use of ISO 32000-1 (PDF/UA-1)*. 2012. http://www.iso.org/iso/catalogue_detail.htm?csnumber=54564

International Standards Organization. *ISO 32000-1:2008 , Document management - Portable document format - Part 1: PDF 1.7*. Geneva, 2008. http://www.iso.org/iso/catalogue_detail.htm?csnumber=51502

McCall, Karen. *Accessible and Usable PDF Documents: Techniques for Document Authors*. First. Paris : Karen McCall, 2005.

McCall, Karen.. *Logical Document Structure Handbook: Word 2003*. Paris: Karen McCall, 2005.

McCall, Karen. *Mapping Section 508 to Digital Document Accessibility*. Karlen Communications. Oakville: Karen McCall, 22 November 2006.

McCall, Karen. "PDF and the User Experience." Karlen Communications, 2016. http://www.karlencommunications.com/PDFsurvey.html

PDF Association. *PDF/UA Competency Centre*. 16 June 2014. http://www.pdfa.org/publication/pdfua-reference-suite/

PDF Association. *PDF/UA in a Nutshell*. Ed. Duff Johnson. 16 June 2013. http://www.pdfa.org/2013/06/pdfua-in-a-nutshell/

PDF Association. *PDF/UA Structure Elements Best Practices 0.1*. 15 12 2015. http://www.pdfa.org/2015/12/announcing-the-structure-elements-best-practice-guide-0-1/

PDF Association. *The Matterhorn Protocol*. 2013. http://www.pdfa.org/wp-content/uploads/2014/06/MatterhornProtocol_1-02.pdf

United States Access Board. "Information and Communications Technology (ICT) Standards and Guidelines." 2015. *United States Access Board*. https://www.access-board.gov/attachments/article/1702/ict-proposed-rule.pdf

Journal on Technology and Persons with Disabilities

ISSN 2330-4219

LIBRARY OF CONGRESS * U.S. ISSN CENTER
ISSN Publisher Liaison Section
Library of Congress
101 Independence Avenue SE
Washington, DC 20540-4284
(202) 707-6452 (voice); (202) 707-6333 (fax)
issn@loc.gov (email); www.loc.gov/issn (web page)